Li DENG, Xin DU, Ji-zhong SHEN. Web page classification based on heterogeneous features and a combination of multiple classifiers. *Frontiers of Information Technology & Electronic Engineering,* 21(7):995-1004.
https://doi.org/10.1631/FITEE.1900240

Web page classification based on heterogeneous features and a combination of multiple classifiers

Key words: Web page classification; Web page features; Combined classifiers

Corresponding author: Ji-zhong SHEN
E-mail: jzshen@zju.edu.cn
ORCID: Ji-zhong SHEN, https://orcid.org/0000-0002-9031-2379

Motivation

- The huge amount of information on the Internet has exploded over time, which provides people access to valuable resources. Web page classification is critical for website management and information retrieval. Therefore, it is necessary to obtain a web page classification method to achieve good classification performance.
- Web page classification based on a single feature is subject to bias. For example, some web pages lack textual information, so it is difficult to accurately classify them based just on textual features. Better classification results can be obtained using other effective features and by fusion of heterogeneous features.
- Moreover, due to the differences in features and classifiers, a sample may be classified incorrectly by one classifier but can be classified correctly by other classifiers. The classification performance can be improved by combining multiple classifiers.

Main idea

- The tree-like structure of HTML tags is exploited to characterize the web page structural features. Heterogeneous textual features and the proposed tree-like structural features are converted into vectors and fused.
- Confidence is proposed here as a criterion to compare the classification results of different classifiers by calculating the classification accuracy of a set of samples. Multiple classifiers are combined to give a good classification result.

Method

1. Vectors are constructed to represent the web page structure of HTML tags. Heterogeneous textual features and the proposed tree-like structural features are converted into vectors and fused.

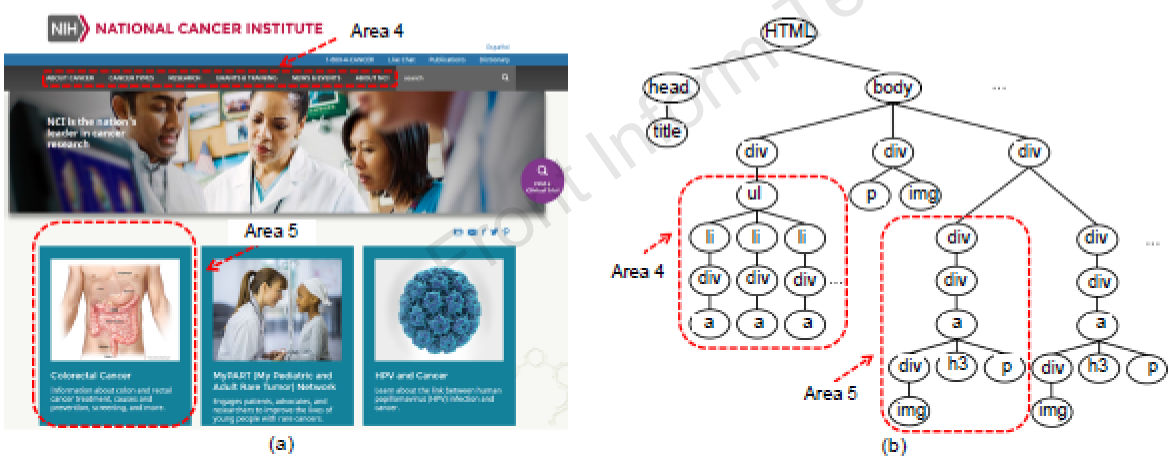

Fig. 4 Web page (a) and tree-like structure (b) of www.cancer.gov

[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
[0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
[0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
[0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
[0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
[0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
[0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
[0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0],
[0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
[0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0],
[0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0].

Method

2. Confidence is proposed here as a criterion to compare the classification results of different classifiers. Then, multiple classifiers are combined to give the final classification result.

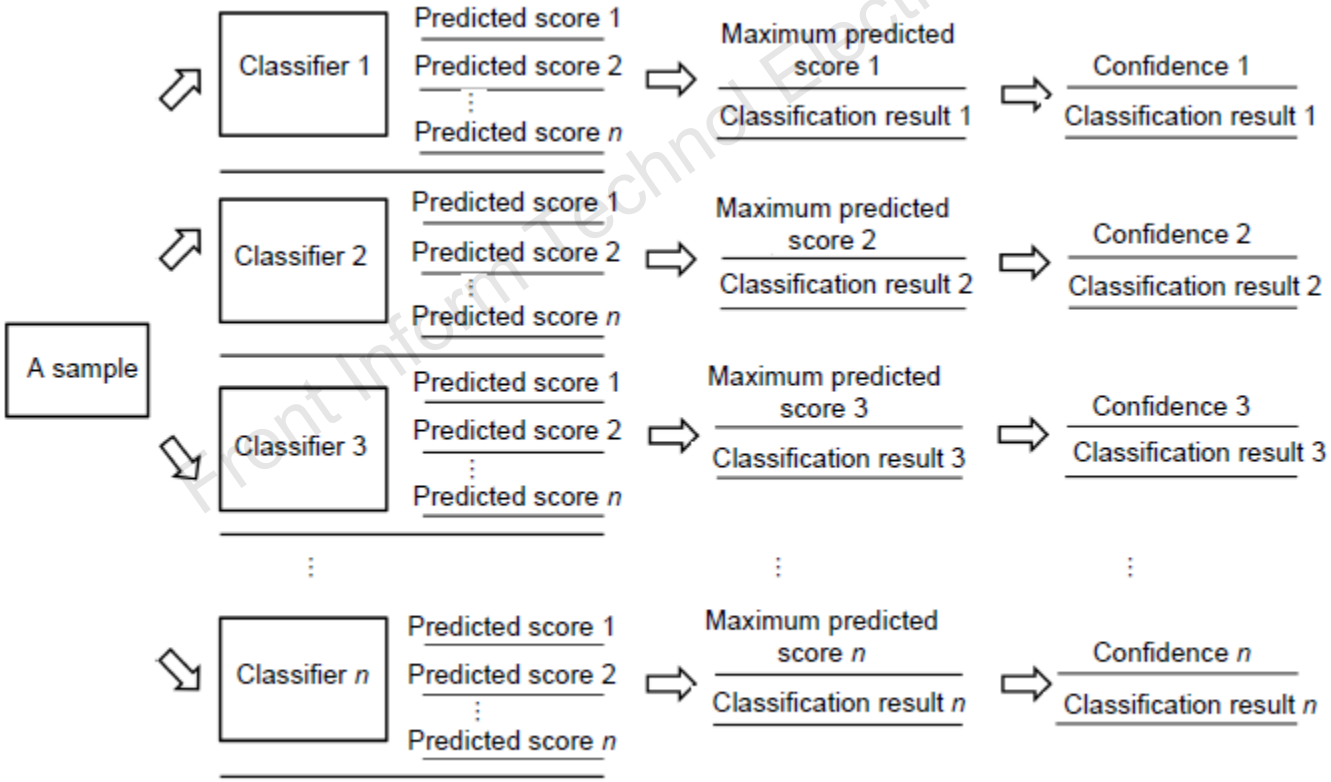

Fig. 6 Process of obtaining classification results and confidences for multiple individual classifiers

Major results

Table 3 Classification accuracy on the Amazon dataset

Method	ACC (%)
LSTM classifier based on textual features	89.2
SVM classifier based on textual features	90.3
LSTM classifier based on fusion of textual and structural features	93.7
SVM classifier based on fusion of textual and structural features	92.4
Combination of LSTM and SVM classifiers based on fusion of textual and structural features	94.2

SVM: support vector machine; LSTM: long short-term memory; ACC: accuracy

Table 5 Comparison results of several works on the 7-web-genres dataset

Method	Precision (%)	Recall (%)	*F*-measure (%)
MCC	92.5	90.0	91.2
RFSE	91.2	89.9	90.5
CSA			91.5
Proposed method	95.5	95.4	95.4

MCC: multi-classifier combination; RFSE: random feature sub-spacing ensemble; CSA: combined stemming approach

Conclusions

- The fusion of textual features and the proposed tree-like structure features is comprehensive and effective.
- The classification results are compared under the same criterion, that is, confidence. Deep neural network and support vector machine classifiers are combined to give a good classification result .
- Experimental results on the Amazon dataset, 7-web-genres dataset, and DMOZ dataset showed that the accuracies are increased to 94.2%, 95.4%, and 95.7%, respectively, by our proposed method, and demonstrated higher accuracy than the related web page classification algorithms.